Stochastic Dynamics and Dominant Protein Folding Pathways

Journal:	*Philosophical Magazine & Philosophical Magazine Letters*
Manuscript ID:	TPHM-08-Jun-0232.R2
Journal Selection:	Philosophical Magazine
Date Submitted by the Author:	07-Nov-2008
Complete List of Authors:	Faccioli, Pietro; University of Trento, Physics Department and INFN; Pietro Faccioli, Pietro Faccioli Sega, Marcello; Frankfurt Institute for Advanced Studies Pederiva, Francesco; Universita' di Trento, Physics Department Orland, Henri; CEA-Saclay, Institut de Physique Theorique
Keywords:	soft matter, statistical physics
Keywords (user supplied):	protein folding
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
Soft2008v3.tex	

Philosophical Magazine
Vol. 00, No. 00, 21 December 2008, 1–6

Conference Proceedings

Stochasitic Dynamics and Dominant Protein Folding Pathways

Pietro Faccioli[a,b] *, Marcello Sega[c], Francesco Pederiva[d], and Henri Orland[e]

[a] *Dipartimento di Fisica, Università degli Studi di Trento, via Sommarive 14, Povo (Trento), 38100 Italy* ; [b] *I.N.F.N., Gruppo Collegato di Trento*; [c] *Frankfurt Institute for Advanced Studies, Ruth-Moufang-Str. 1, D-60435 Frankfurt, Germany*; [d] *Institut de Physique Théorique, Centre d'Etudes de Saclay, CEA, IPhT, F-91191, Gif-sur-Yvette, France.*;

(*Received 00 Month 200x; final version received 00 Month 200x*)

Keywords: Polymer Dynamics, Protein Folding, Stochastic Path Integrals

We present the results of a recently developed theoretical framework denominated Dominant Reaction Pathways (DRP), to study thermally activated reactions in multi-dimensional systems. In particular, we focus on application to the protein folding reaction. By applying the saddle-point approximation to the stochastic path integral generated by the Langevin Equation, we derive a least-action principle which allows to rigorously determine directly the most probable reaction pathways, bypassing the long-standing computational problems associated with the decoupling of time-scales in the problem. We show the results of a number validation studies in which the accuracy of the DRP approach was assessed studying molecular transitions. In all cases, the DRP predictions are found to be consistent with the MD results, but extremely less computationally expensive.

1. Introduction

The problem of characterizing thermally activated transitions of bio-molecules is crucial for the understanding of many phenomena at the interface of physics, chemistry and biology. In particular, the so-called "Part II" of the Protein Folding Problem consists in identifying the sequence of conformational transitions which proteins assume on their way from the denatured state to the native state ("Part I" being understanding the sequence structure relationship).

From a theoretical perspective, the most natural strategy to approach this problem would be to integrate numerically the equations of motion, e.g. via Molecular Dynamics (MD) simulations. In fact, from a statistically significant sample of MD transitions it is possible to identify the transition state and compute folding rates, which can be matched against experimental data.

Unfortunately, state-of-the-art all-atom MD simulations cover time intervals in the range $10 - 10^2\ ns$ (see e.g. [1] and references therein) while folding transitions take place in the $ms - s$ time range. Hence, MD can be useful only in investigating the reactions leading to local secondary structures formation, i.e. α-helices and β-sheets. The reason for such a decoupling of time scales is that the folding transitions

*Corresponding author. Email: faccioli@science.unitn.it

ISSN: 1478-6435 print/ISSN 1478-6443 online

DOI: 10.1080/1478643YYxxxxxxxx
http://www.informaworld.com

involve overcoming free-energy barriers of several units of $k_B\ T$. As result, the computational time is entirely invested in describing the motion of the molecule in the molten globule state (for folding reactions) or in the native state (for unfolding reactions) . On the other hand, one is interested in the information encoded in the reaction pathways joining these two states.

Given the limitations of the available MD simulations, several alternative approaches have been recently developed, to deal with the problem of characterizing the entire reaction in a high-dimensional space, using available computers. A drawback of these methods is that they rely either on *ad-hoc* assumptions about the dynamics obeyed by the system [2] or on a specific choice of reaction coordinates [3].

In this contribution to the Conference Proceedings, we shall present the results of the DRP method [4–7], which allows to rigorously identify the statistically significant protein folding pathways, without relying on any choice of reaction coordinates and without involving any uncontrolled *ad hoc* approximation.

The major advantage of the DRP approach is that it avoids investing computational time in simulating the local thermal motion in the metastable configurations. This is possible because the key equation can be formulated in a form which does not depend explicitly on the time variable. As a result, the computational difficulties associated with the existence of different time scales in rare thermally activated reactions, are naturally and rigorously bypassed and studying conformational transitions of large molecules in atomistic details becomes feasible on available computers.

2. Brief review of the DRP Approach

The starting point of the DRP approach is the path integral representation of the Fokker-Planck conditional probability to perform a transition from an initial configuration x_i to a final configuration x_f in a time t, i.e.

$$P(x_f,t|x_i,0) = e^{-\frac{(U(x_f)-U(x_i))}{2k_BT}} \int_{x_i}^{x_f} \mathcal{D}x(\tau)\ e^{-\int_0^t d\tau\ \left(\frac{\dot{x}^2(\tau)}{4D}+V_{eff}[x(\tau)]\right)}. \qquad (1)$$

D is a constant diffusion coefficient, and $V_{eff}(x)$ is an effective potential defined as

$$V_{eff}(x) = \frac{D}{4(k_BT)^2}\left((\nabla U(x))^2 - 2k_BT\nabla^2 U(x)\right), \qquad (2)$$

$x = (\mathbf{x}_1, ..., \mathbf{x}_{N_p})$ is a vector specifying the position of all the N_p constituents of the molecule (atoms or amino-acids). In the specific case of the study of the protein folding reaction, the initial configuration x_i may be taken in the denatured state, while the final configuration $x_f \equiv x_N$ lies in native state. The representation (1) is analog to a quantum mechanical path integral in imaginary time, in a system specified by the action $S_{eff}[x] = \int_0^t d\tau\ \left(\frac{\dot{x}^2(\tau)}{4D} + V_{eff}[x(\tau)]\right)$.

The most probable paths contributing to (1) are those for which the exponential weight $e^{-S_{eff}}$ is maximum, hence for which S_{eff} is minimum. A trajectory which connects configurations that are not classically accessible in the absence of thermal fluctuations corresponds to an instanton in the quantum-mechanical language. In the context of diffusive dynamics we shall refer to these as to the *dominant reaction pathways*. Determining these trajectories for realistic proteins using con-

ventional methods —such as Molecular Dynamics— is extremely challenging from the computational point of view. In addition to the numerical difficulties associated with the existence of very different time scales, one has also to face the solution of boundary-value problems, which are considerably harder than initial-value problems.

Fortunately, a dramatic simplification is obtained upon observing that the dynamics described by the effective action S_{eff} is energy-conserving and time-reversible. This property allows us to switch from the *time*-dependent Newtonian description to the *energy*-dependent Hamilton-Jacobi (HJ) description. We note that this could not be done at the level of the Langevin equations. In the HJ framework, the dominant pathways connecting given initial and final positions is obtained by minimizing numerically —e.g. via simulated annealing– a discretized version of the target function (HJ functional)

$$S_{HJ} = \int_{x_i}^{x_f} dl \sqrt{\frac{1}{D}\left(E_{eff} + V_{eff}[x(l)]\right)}, \tag{3}$$

where dl is an infinitesimal displacement along the path trajectory. E_{eff} is a free parameter which determines the total time elapsed during the transition, according to:

$$t_f - t_i = \int_{x_i}^{x_f} dl \sqrt{\frac{1}{4D\left(E_{eff} + V_{eff}[x(l)]\right)}}. \tag{4}$$

In [5, 6] we showed that for thermally activated transitions one needs to choose $E_{eff} = -V_{eff}(x_N)$.

The HJ formulation of the dynamics leads to an impressive computational simplification of this problem. In fact, the total Euclidean distance between the coil state and the native state of a typical protein is only 1-2 orders of magnitude larger than the most microscopic length scale, i.e. the typical monomer (or atom) size. As a consequence, only $\sim 30 - 50$ discretized displacement steps are usually sufficient for convergence. This number should be compared with 10^{12} time-steps required in the time-dependent Newtonian description. The physical reason why the HJ formulation is so much more efficient compared to the Newtonian formulation is the following: in traditional Molecular Dynamics simulations, proteins spend most of their time in meta-stable minima, trying to overcome free-energy barriers. The HJ formulation avoids investing computational times in such "waiting" phases by considering intervals of fixed displacements, rather than fixed time-length.

Once the dominant pathways have been determined, it is also possible to systematically account for the effects of quadratic thermal fluctuations around them, by means of the Monte Carlo algorithm [6]. Let us denote by

$$\overline{x}(n) = (\overline{\mathbf{x}}_1(n), ..., \overline{\mathbf{x}}_{N_p}(n)), \tag{5}$$

with $n = 1, .., N$, the dominant pathway trajectory corresponding to a sequence of configurations of the molecule, and determined by minimizing numerically the HJ action (3). The time at which each configuration $\overline{x}(n)$ is visited during the transition can be obtained by computing the set of time intervals separating each of the path steps $\overline{x}(n)$ from $\overline{x}(n+1)$:

$$\Delta\tau_{n,n+1} = \frac{\Delta l_{n,n+1}}{\sqrt{4D(V_{eff}(\overline{x}(n)) - V_{eff}(\overline{x}(1)))}}. \tag{6}$$

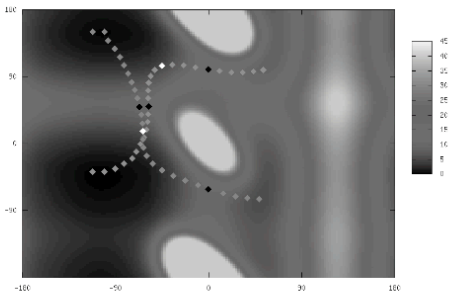


Figure 1. Dominant Reaction Paths for the $C7_{ax} \rightarrow C7_{eq}$ (red squares, from bottom right to top left) and $\alpha_L \rightarrow \alpha_R$ (blue squares, from bottom left to top right) transitions. In the background, the free-energy prole for the and ϕ and ψ dihedrals is shown (in units of kJ=mol). Black and white squares identify the minimum residence time conformations and the commitment analysis transition states, respectively.

The sequence of time intervals obtained this way can be used to write a discretized version of the transition probability (1):

$$P(x_f, x_i; \tau(N)) = \int \prod_{n=1}^{N-1} \left[d\mathbf{x}_1(n) ... d\mathbf{x}_{N_p}(n) \right] e^{-\sum_{n=1}^{N-1} \Delta\tau_{n,n+1} \left[\frac{1}{4D} \sum_{i=1}^{N_p} \left(\frac{\mathbf{x}_i(n+1) - \mathbf{x}_i(n)}{\Delta\tau_{n,n+1}} \right)^2 + V_{eff}(\mathbf{x}_1(n), ..., \mathbf{x}_{N_p}(n)) \right]} \quad (7)$$

We stress the fact that in Eq. (7) the time intervals are chosen large (small), when the most probable trajectory at that time is evolving slowly (fast). In other words, the information encoded in the dominant pathway has been used to identify a particularly convenient discretized representation of the path integral (1), in which the sizes of the time steps are adapted according to the evolution of the most probable pathways.

3. Assessing the Reliability of the DRP Approach

The first validation study of the DRP approach was performed in [5], in which we studied the kinetics of alanine dipeptide, using the *GROMOS96* force field.

In Fig. 1, we present the results of the analysis relative to two specic transitions ($C7_{ax} \rightarrow C7_{eq}$ and $\alpha_L \rightarrow \alpha_R$), compared with the free-energy landscape computed by direct integration. The values of the two ϕ and ψ dihedrals along the paths obtained by minimizing the effective action are plotted on top of the relative free-energy map. These simulations were performed at temperature $T = 300K$, and assuming a diffusion constant of $D = 0.02A^2ps^{-1}$ for all atoms. An analysis of the residence time along the path shows that in each of the two dominant pathways, there are two points where the conformation of alanine dipeptide has shortest residence time. These points, indicated in Fig. 1 with black symbols, very accurately locate the maxima of the free energy, along the path. In all similar analysis performed on this system, we observed that the information encoded in the dominant pathway trajectories was in excellent agreement with the results of MD simulations, performed in the same model. On the other hand, the computational gain of adopting the DRP method was impressive: the characterization of the relevant transitions in atomistic detail took just few minutes on a regular desktop, while

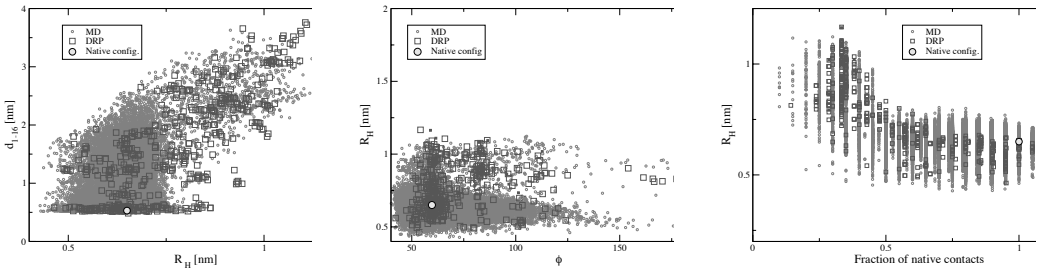


Figure 2. Comparison between MD and DRP calculations for three pairs of order parameters (d_{1-16} vs R_H, R_H vs ϕ and R_H vs N_c. Circles are configurations obtained via MD simulations, squares are configurations obtained from the DRP analysis. The gray circle is the value of the order parameters in the native configuration.

about a week of calculations on the same machine was required, in order to extract the same amount of information, from the standard MD techniques.

The next step in our validation study consists investigating the reliability of the DRP approach, when it is applied to characterize the folding reaction of a longer polypeptide chain. Since we are interested in comparing the results of the DRP method with those of traditional MD simulations, we chose a relatively small chain and adopted a coarse-grained model: we studied the folding of 16-residue β-hairpin GB1, using a Go-type force field, between the amino-acids. In fact, the dynamics of the Go model is sufficiently simple to allow several folding-unfolding trajectories to be generated by MD on a regular workstation, at an affordable computational cost. On the other hand, such a model displays some of the complexity which we expect to face, in more realistic and detailed all-atom simulations, of long polypeptide chains. For example, the molecule presents a unique stable fold and a ensemble of denatured configurations with large conformational entropy.

In the three panels of Fig.2 we present some preliminary results in which the predictions of the MD and DRP approaches are compared. In order to analyze the results we project the dynamics on the three planes selected by the following set of order parameters: The distance d_{1-16} between the C and N terminus of the chain, the angle ϕ at the turn of the hairpin, i.e. between the ASP-ALA and LYS residues, the fraction of native contacts[1] and the radius of gyration R_H of the group defined by the hydrophobic residues Phe, Tyr, and Trp.

The points correspond to configurations obtained by running 5 independent MD trajectories, of $15ns$ each. The squares represents points obtained in the DRP approach, starting from 10 indipenent dominant pathways and sampling 3 different thermal fluctuations around each of them. The gray circle represents the value of the order parameters in the native state conformation, which was downloaded from the Protein Data Bank.

We can see that the DRP method correctly identifies the regions of configuration space which are explored by the MD trajectories. This fact is observed in all the combinations of the selected set of reaction coordinates. More quantitative comparison between MD and DRP, based on the study of the free energy landscape is in progress [7]. It is important to stress that we do *not* expect the configurations vis-

[1]Two residues are considered in contact if their distance is less than 6 A. The fraction of native contacts in Fig. 2 is defined as the number of non-consecutive residues in contact, divided by the number of non-consecutive residues in contact in the native state.

ited by the DRP trajectories to accumulate in the regions where the density of MD configurations is highest. In fact, by definition, the density of MD configurations increases in the vicinity of the (meta-)stable states and becomes small (but finite) in the vicinity of the configurations visited during the folding-unfolding transitions. On the other hand, by construction, each dominant pathway trajectory leaves the native state and visits denatured conformations, by steps of equal $dl = \sqrt{\sum_i dx_i^2}$ and do not accumulate in the vicinity of stable and meta-stable states.

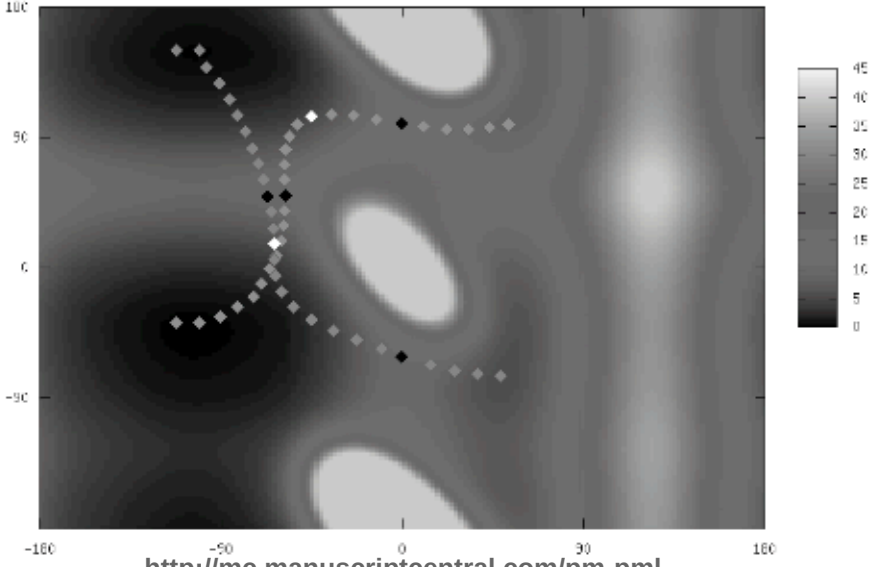
The analysis of the dominant pathway trajectories can be used to investigate the folding mechanism oft the β-hairpin, within the present simple Go-type model. We have found support for a hydrophobic collapse mechanism. [7]. In fact, in the initial (final) stage of the unfolding (folding) reaction, the size of the hydrophobic core remains constant and consistent with the value in the native state, while the number of native contacts rapidly drops (increases).

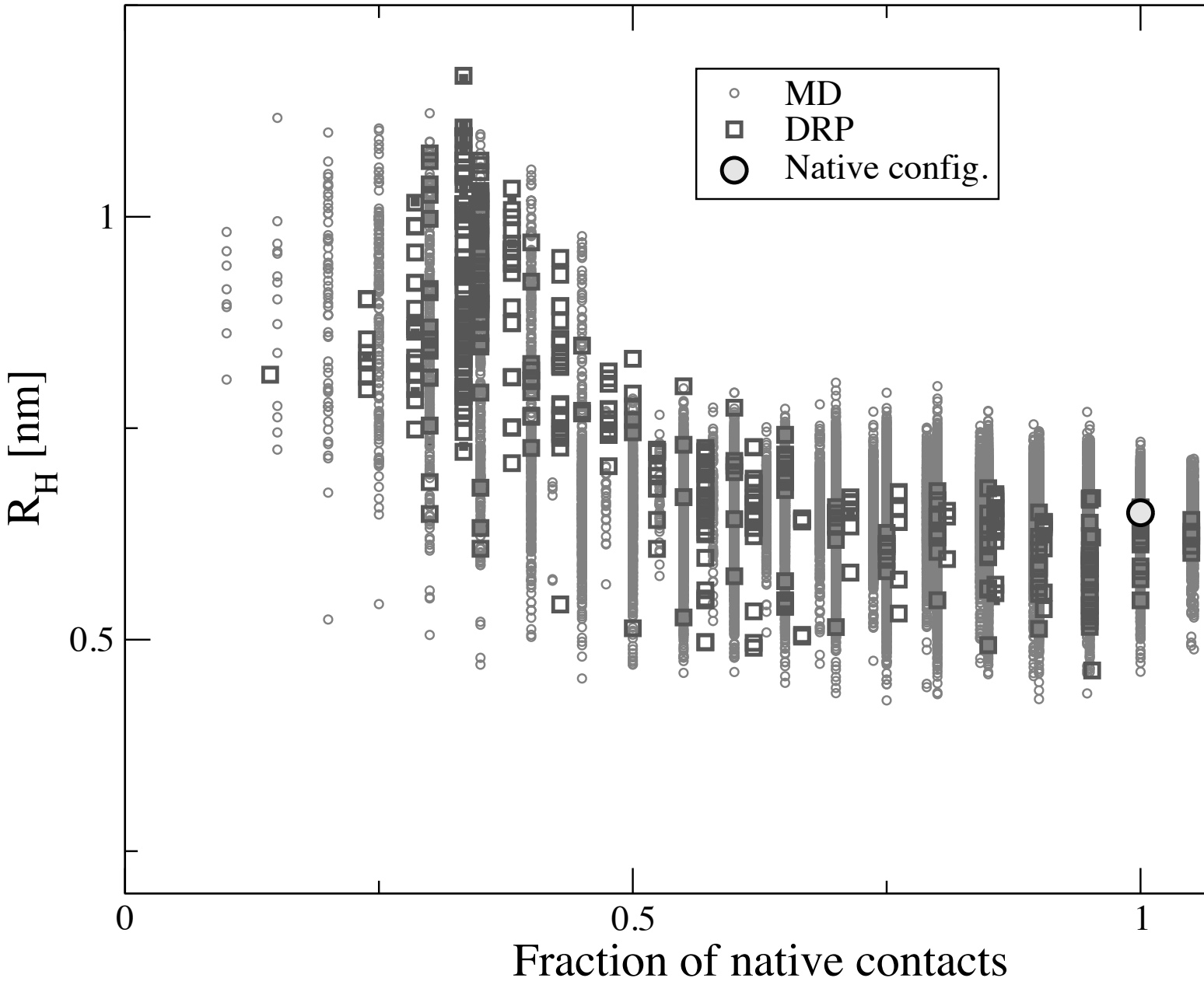
4. Conclusions and future developments

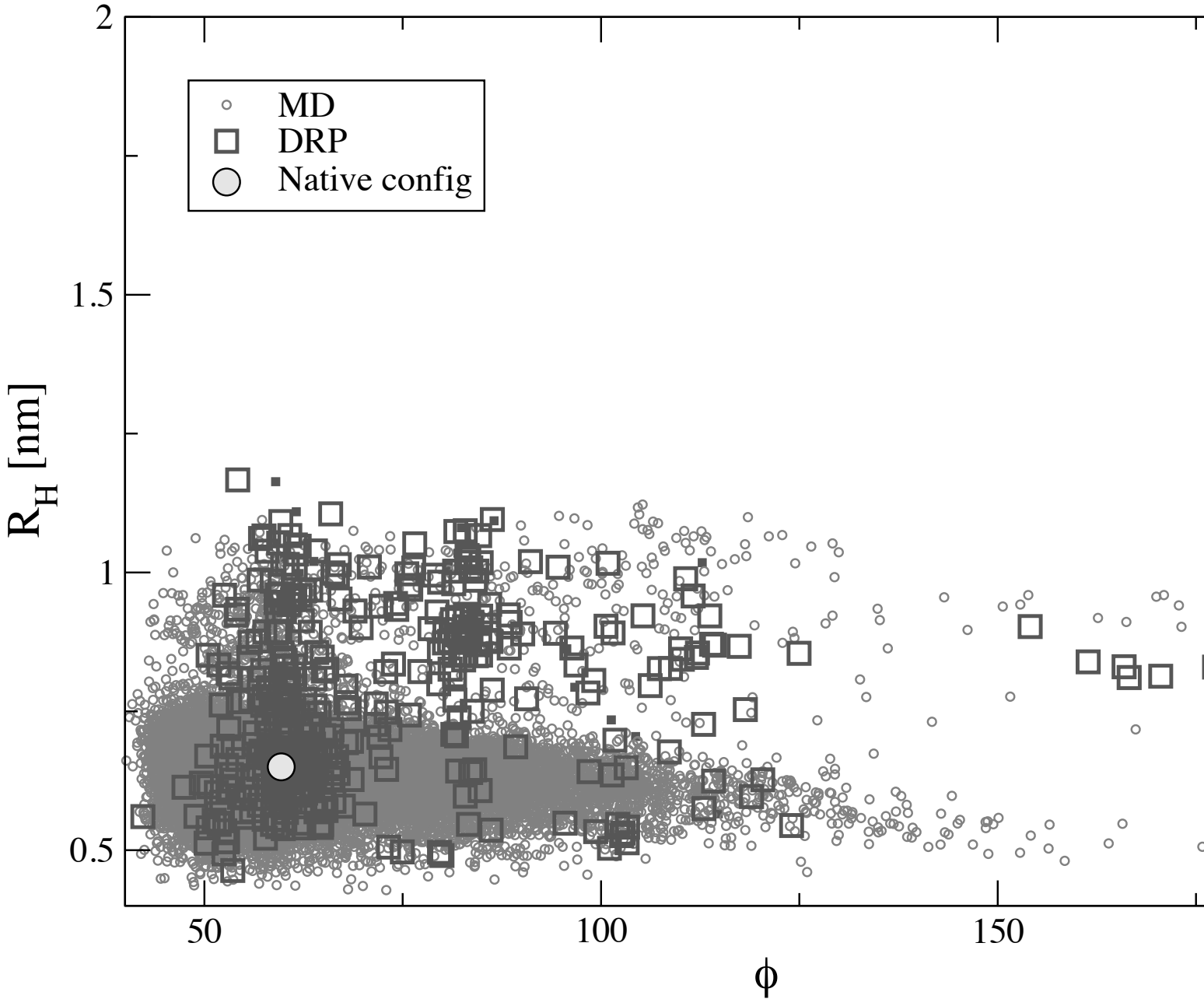
In this proceedings, we have presented some recent applications of the DRP method to study conformational transitions of macromolecules. In the analysis of transitions of alanine-dipeptide using the *GROMOS94* force field, and in the study of the folding of the 16-residue β-hairpin terminus of protein GB1 using a coarse-grained model, the DRP predictions were confronted with the results of MD simulations. We have found that the two methods lead to consistent pictures of the transition. On the other hand, the DRP approach leads to a huge computational gain, relative to traditional MD simulations. This development opens the door to a characterization of thermally activated reactions involving large molecules such as proteins, in atomistic detail using available computers.

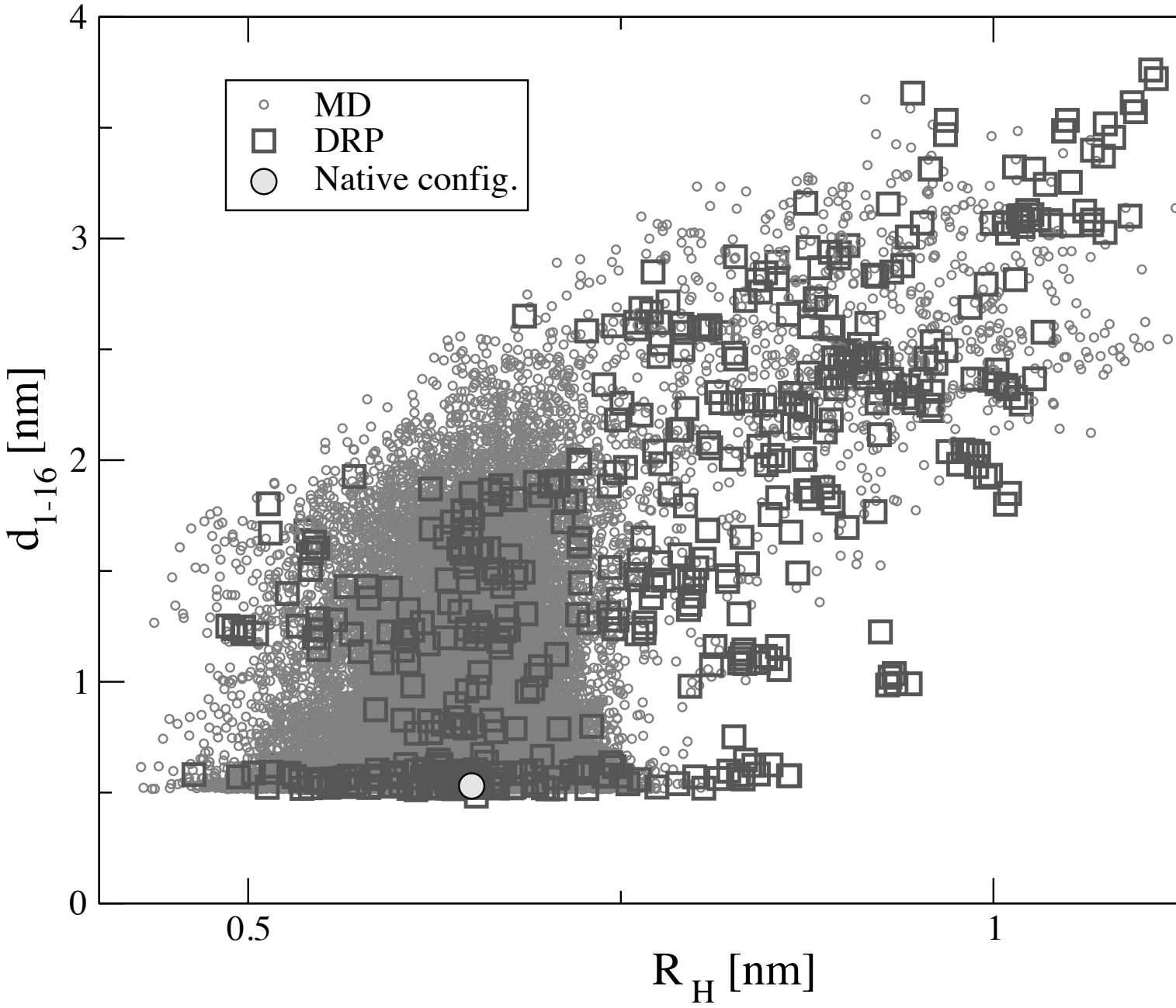
References

[1] V. S. Pande et al., Biopolym. 68 (2003) p. 91.
[2] R. Olender, R. Elber, J. Chem. Phys. 105 (1996) p. 9299; R. Elber, A. Ghosh and A. Cardenas, Acc. Chem. Res 35, (2002) p. 396. A. Ghosh, R. Elber and H.A. Sheraga, PNAS 99 (2002) p. 0394.
[3] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Ann. Rev. Phys. Chem. 53, (2002) p. 291. P.G. Bolhuis *et. al.*, Ann. Rev. Phys. Chem. 53 (2002) p. 291;
[4] P.Faccioli, M.Sega, F.Pederiva and H.Orland, Phys. Rev. Lett. 97 (2006) p. 108101.
[5] M.Sega, P.Faccioli, G.Garberoglio, F.Pederiva and H.Orland, Phys. Rev. Lett. 99 (2007), p. 118102.
[6] E.Autieri, P.Faccioli, M.Sega, F.Pederiva and H.Orland, xArchiv: 0806.0236 (cond-mat)
[7] P.Faccioli, arXiv:0806.3734, J. Chem. Phys. , in press.
[8] M.M. Klosek, B.J. Matkowsky and Z. Schuss and Ber. Bunsenges. Phys. Chem. 95, (1991) p. 331.
[9] Ensign DL, Kasson PM, Pande VS R. Du, J. Mol. Biol. 374 (2007) p. 806. V.S. Pande, A.Y. Grosberg,T. Tanaka and E.S. Shakhnovich, J. Chem. Phys **108** (1998) 334
[10] V. Munoz *et al.* Nature 390 (1997), p. 196. V. Munoz, *et al.* Proc. Natl. Acad.Sci. USA 95 (1998).


100
50
0
-50
-100
-50
0
50
100
45
40
35
30
25
20
15
10
5
0


MD
DRP
Native config.
R_H [nm]
1
0.5
0
0.5
1
Fraction of native contacts


MD
DRP
Native config
2
1.5
1
0.5
R_H [nm]
50
100
150
ϕ


MD
DRP
Native config.
$d_{1\text{-}16}$ [nm]
R_H [nm]
0
1
2
3
4
0.5
1